INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Projet CARAVEL

Systèmes de médiation d'information

Rocquencourt

THÈME 3A

Table des matières

1 Composition de l'équipe

Responsable scientifique
Eric Simon [DR Inria]
Responsable permanent
François Llirbat [CR]
Assistante de projet
Elisabeth Baqué [AI]
Collaborateurs extérieurs
Luc Bouganim [MC, université de Versailles]
Mokrane Bouzeghoub [Professeur, université de Versailles]
Chercheur invité
Judy Cushing [Oregon Graduate Institute, USA, 5 mois]
Ingénieurs experts
Mokrane Amzal
Françoise Fabret
Doctorants
Helena Galhardas [université de Lisbonne, jusqu'au 25 septembre]
Alberto Lerner [université de Rio de Janeiro]
Ioana Manolescu [université de Versailles]
Joao Pereira [université de Lisbonne, jusqu'au 25 septembre]
Khaled Yagoub [boursier MESR, université de Versailles, jusqu'au 31 Mai]
Stagiaires
Jean-Pierre Matsumoto [Université Paris VI, stage de DEA]
Aurelian Lavric [Université Paris VI, stage de DEA]
Cristian-Augustin Saita [Université Versailles, stage de DEA]
Lucian Precup [Ecole Polytechnique de Bucarest]
Gabriel Kaltman [Ecole Polytechnique de Bucarest]
Alexandru Carstoiu [Ecole Polytechnique de Bucarest]

2 Présentation et objectifs généraux

L'énorme quantité d'informations aujourd'hui disponible sur le Web et la très grande disparité de ces informations aussi bien dans leur contenu que dans leur mode d'accès, rendent bien souvent laborieuse la recherche de données précises. Chacun aimerait avoir accès à une vue intégrée et à jour de ces informations, ce qui suppose à la fois une structuration uniforme et cohérente des sources d'information et des modes d'interrogation adaptés aux besoins de l'utilisateur. Le projet Caravel répond à ce problème fondamental d'intégration de sources d'informations au travers de trois grands thèmes de recherche complémentaires :

- Thème 1 : il s'agit de faciliter la publication de ressources dans un réseau ainsi que l'accès à ces ressources au moyen de langages de haut niveau. Les ressources peuvent être des données (structurées ou non) ou des services (librairies, programmes scientifiques, sites Web, etc), l'ensemble formant un *système d'information global*. Deux difficultés majeures se posent : réduire considérablement l'effort de développement nécessaire à la publication

de ressources et mettre au point des méthodes d'optimisation pour les langages de haut niveau proposés.

- Thème 2 : il s'agit de faciliter la production de données élaborées à partir de données et de services publiés dans le système d'information global. Les principales difficultés à résoudre sont d'une part d'intégrer des données hétérogènes de façon cohérente, correcte et efficace, et, d'autre part d'assembler judicieusement des programmes disparates dans une chaîne de traitement de données et enfin d'exécuter efficacement de telles chaînes de traitement.
- Thème 3 : il s'agit de faciliter la consultation de données « pour tous », c'est à dire sans formuler de requêtes dans un langage de requêtes de bases de données. La consultation de données via un site Web est une première solution, mais les difficultés sont d'administrer la structure logique d'un site tout en garantissant de bonnes performances de consultation et d'offrir des méthodes de navigation adaptatives en fonction de l'intérêt de l'utilisateur.

Plusieurs actions de recherche sont menées dans chacun de ces thèmes. Deux grandes actions structurent le premier thème. La première s'inscrit dans une approche de type « pull » pour l'accès aux ressources du système d'information global, tandis que la seconde se situe dans une approche de type « push ». Le second thème distingue également deux actions de recherche qui visent à aider d'une part à la génération de programmes efficaces de nettoyage de données hétérogènes (« data cleaning », en anglais) et d'autre part à l'assemblage et l'exécution de workflow scientifiques distribués. Enfin, le dernier thème recouvre deux actions qui visent à faciliter l'administration de sites Web performants et à offrir des moyens de navigation adaptatifs à l'utilisateur.

Les techniques conçues dans ces actions de recherche prennent la forme de langages de bases de données, de modèles de données ou d'algorithmes. Ces techniques sont implantées dans des composants logiciels modulaires qui s'interfacent entre des applications clientes et des serveurs d'information selon un modèle d'architecture à trois-tiers. D'un point de vue stratégique, nous concevons des composants logiciels facilement assemblables entre eux, ce qui facilite leur utilisation combinée dans le déploiement d'applications et permet une grande synergie entre les différentes actions de recherche du projet. De plus, nous nous efforçons d'expérimenter nos composants dans le cadre d'applications réelles en collaboration avec des partenaires utilisateur via des contrats industriels.

3 Fondements scientifiques

L'histoire de la recherche en bases de données est exceptionnelle par sa productivité, son transfert industriel et son impact économique. Reconnue depuis un peu plus de 20 ans comme une discipline de recherche de base par les Etats-Unis suivis par la plupart des pays industrialisés, la recherche en bases de données a été conduite d'abord dans les laboratoires des grands groupes industriels pour être généralisée ensuite aux laboratoires publics et universités. Les Systèmes de Gestion de Bases de Données (SGBD) sont aujourd'hui des logiciels de base essentiels dans tout système d'information. Intuitivement, un SGBD permet à des utilisateurs de poser avec une certaine souplesse des requêtes pour manipuler (rechercher et modifier) une grande masse de données persistantes. Il doit contrôler la concurrence de ces accès tout en

garantissant la cohérence, l'intégrité, la confidentialité et la sécurité des données.

Depuis l'apparition vers la fin des années 60 des premiers SGBD hérités des systèmes de gestion de fichiers, d'importants résultats théoriques et pratiques ont ponctué l'histoire des bases de données. L'invention du *modèle relationnel* en 1970 est l'événement le plus marquant (il a valu à son auteur, Tedd Codd, le prix Turing de l'ACM en 1982). Le modèle relationnel a permis d'établir les fondements mathématiques qui manquaient au domaine et a ouvert de grandes perspectives de recherche, notamment en conception de schémas normalisés et en langages de requêtes déclaratifs. Les premières retombées de ces recherches ont été de faciliter l'administration et la manipulation de bases de données et d'accroître la productivité des utilisateurs.

Cependant, la puissance des langages relationnels qui permettent d'exprimer des requêtes complexes a longtemps posé des problèmes de performances. Ceux-ci ont été progressivement résolus par des efforts continus, durant plus de quinze ans, en recherche et développement, avec en particulier des algorithmes efficaces pour traiter les opérateurs relationnels, des techniques d'optimisation de requêtes, le support intégré efficace des transactions et l'exploitation du parallélisme pour exécuter les opérateurs relationnels sur calculateur multiprocesseur. Ces deux derniers points ont valu a Jim Gray le prix Turing de l'ACM en 1998.

Depuis 1981, les projets Sabre puis Rodin ont participé activement à ce mouvement de la recherche en concevant et en expérimentant des techniques afin d'améliorer les fonctionnalités et les performances des SGBD. Ces techniques ont pris la forme de langages et de modèles à base de règles et d'objets qui étendent la puissance d'expression des modèles de bases de données existant, d'algorithmes d'optimisation de langages de bases de données ainsi que d'algorithmes et de structures de données pour l'exécution d'opérations coûteuses de bases de données et pour l'exécution concurrente de transactions. Diverses collaborations avec des industriels (surtout via des contrats européens) nous ont permis d'évaluer nos solutions dans des systèmes complets impossibles à développer dans le contexte d'un projet Inria (e.g., évaluation de nos algorithmes d'optimisation de requêtes pour bases de données parallèles sur le système DBS3 de Bull sur machine KSR, ou évaluation d'un protocole de contrôle de concurrence dans le système Validity developpé par la société NCM). Enfin, ces collaborations ont donné lieu à des transferts industriels de logiciels (e.g., Omnis, Disco) ou de solutions integrées à des produits (e.g., O2 Engine, Java Universal Binding).

Avec la création du projet Caravel, nous avons redéfini notre problématique de recherche autour de l'intégration d'information dans un réseau composé de sources d'information hétérogènes et autonomes. Deux raisons principales fondent nos décisions. La première est l'évolution des applications de base de données résultant des progrès technologiques, de l'explosion du Web et de l'internet, ainsi que de l'importance croissante des applications d'aide à la décision. La deuxième raison est le degré de maturité auquel sont parvenus les SGBD commercialisés. Ce dernier point a deux conséquences : d'une part certains problèmes sont maintenant considérés comme résolus et d'autre part, les industriels sont souvent les mieux placés pour continuer à améliorer les performances et les fonctionalités des noyaux de SGBD. Nos recherches actuelles s'appuient considérablement sur notre expérience en conception de langages et de modèles de bases de données ainsi qu'en algorithmes d'exécution distribuée d'opérations de bases de données et d'optimisation de langages.

4 Domaines d'applications

La stratégie du projet Caravel repose sur des collaborations avec des partenaires utilisateurs dans des contrats de recherche à finalité applicative. Cette stratégie nous permet de mieux comprendre les besoins d'applications complexes dans le domaine de l'intégration d'information et d'identifier des problèmes de recherche nouveaux (e.g., modèle de workflow scientifiques, modèle et algorithmes pour le nettoyage de données, middleware pour les applications scientifiques). De plus, la collaboration avec des utilisateurs nous offre les moyens d'expérimenter nos solutions dans des contextes d'utilisation réelle.

Jusqu'à présent, le projet s'est surtout intéressé aux systèmes d'information pour l'environnement car c'est un domaine d'application très riche en problèmes d'intégration d'information : les problèmes d'intégration se posent à grande échelle, les sources d'information ont une forte autonomie due à la pluridisciplinarité du domaine et l'intégration d'information est nécessaire aux nombreuses applications d'aide à la décision. Le choix d'un domaine d'application nous a permis d'accumuler depuis cinq ans une expertise reconnue, ce qui facilite le développement de nouvelles collaborations et l'approfondissement des problèmes de recherche qui nous concernent. Les applications principales sur lesquelles nous travaillons sont la gestion de ressources naturelles en zones côtières (projet européen Thetis), la prédiction de la qualité de l'air en milieu urbain (projet européen Decair) et l'analyse des phénomènes de bio-corrosion sur les plates-formes pétrolières (projet Ecobase).

D'autres applications environnementales sont en cours d'exploration comme la gestion de risques liés à des phénomènes naturels (par exemple, inondations) ou à des accidents (par exemple, marées noires). Mais depuis cette année, nous examinons aussi des applications dans le domaine de la santé qui présentent des caractéristiques semblables aux applications que nous avons déjà étudiées : dossier électronique du patient, base de données génétiques universelle (avec le Centre National de Génotypage), gestion de données en neuroimagerie (projet d'action de recherche coopérative).

5 Logiciels

Cette année, nous avons poursuivi notre effort sur le développement de composants logiciels qui intègrent des solutions élaborées au cours des années précédentes. Cinq prototypes de composants logiciels ont déjà été démontrés l'année dernière au cours des deux plus importantes conférences internationales en bases de données, SIGMOD et VLDB (ces démonstrations sont sélectionnées par un comité d'évaluation). Un point important est la mise au point de méthodes de développement et l'utilisation d'outils de génie logiciel destinés à améliorer la robustesse et la pérénité de nos logiciels. Enfin, deux logiciels sont actuellement utilisés en dehors du projet : Le Select et Weave.

La figure ci-dessous donne un synopsis des composants logiciels développés dans le projet Caravel et de leurs interactions potentielles. Les logiciels Le Select/Agora et Le Subscribe sont deux réponses possibles au problème abordé dans le thème 1 du projet. Chacun de ces logiciels offre une vue uniforme et intégrée des informations disponibles dans un système global, mais à travers des modes d'accès différents : Le Select suit une approche de type « pull » (i.e., requête/réponses) et Le Subscribe suit une approche de type « push » (i.e., abonne-

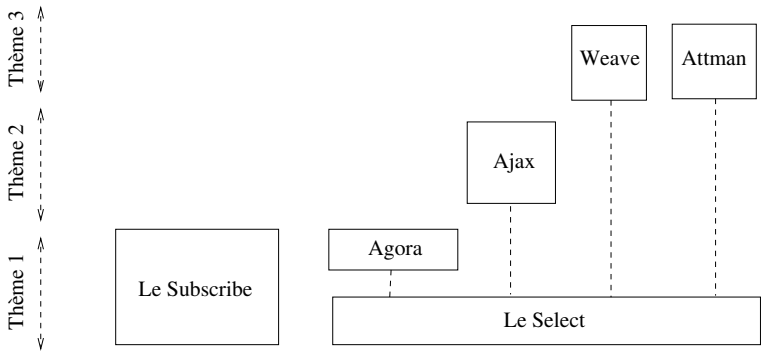

FIG. 1 – Vue d'ensemble des logiciels

ment/notification). Le logiciel Agora donne une représentation uniforme en XML des données tandis que Le Select en donne une vue relationnelle étendue. Le logiciel Ajax qui s'inscrit dans le thème 2 du projet, aide à la génération de programmes efficaces de nettoyage de données extraites par exemple du système d'information global via Le Select. Les données produites par Ajax peuvent à leur tour être publiées via Le Select. Les logiciels Attman et Weave s'inscrivent dans le thème 3 du projet. Ils permettent de construire des sites Web à partir de données qui seraient accédées via Le Select.

5.1 Le Select

Participants : Mokrane Amzal, Ioana Manolescu, Eric Simon [correspondant], Aurelian Lavric, Jean-Pierre Matsumo, Gabriel Kaltman, Lucian Precup.

Le Select est un nouveau système middleware développé depuis 1998 dans le cadre du projet européen Thetis pour répondre aux besoins des applications scientifiques de partager des données et des programmes. Le Select est un successeur du système Disco (développé dans le projet Rodin de 1995 à 1998 dans le cadre de l'action Dyade Médiation, puis transféré en 1999 à la société LibertyMarket qui commercialise le portail Kelkoo.com). Le Select possède une architecture distribuée « peer to peer » de type médiateur/adaptateur. L'objectif général de Le Select est de permettre à des auteurs de ressources, c'est à dire de données ou de services, de facilement publier ces ressources vers une communauté d'utilisateurs en leur donnant une vue uniforme et intégrée et enfin de permettre à des utilisateurs de manipuler cette vue uniforme à travers un langage de haut niveau. Les données ont une représentation uniforme exprimée dans le modèle de données relationnel étendu à des types de données définis par l'utilisateur. La première version de ce logiciel est diffusé depuis le mois d'Octobre. Le Select est actuellement utilisé par plusieurs universités (UNIRIO, UFRJ, IME et PUC-Rio au Brésil), centres de recherche (CNR en Italie, CEMAGREF en France, ICS-FORTH et IMBC en Grèce) et sociétés (Alcatel Industries en France, HR-Wallingford en Angleterre) pour le développement d'applications environnementales. Deux applications ont déja été démontrées dans le cadre du projet Thetis. Cette année, nous avons organisé un seminaire de formation de trois jours destiné aux utilisateurs de Le Select dans divers organismes et universités français qui a rassemblé

dix personnes. De plus, trois nouvelles universités utilisent Le Select a des fins de recherche en médiation d'information : Oregon Graduate Institute (Portland, USA), Univ of Arizona (Tucson, USA) et Univ. de Recife (Recife, Bresil).

5.2 Agora

Participant : Ioana Manolescu [correspondant].

Le système Agora offre une vue uniforme en XML des sources de données publiées à l'aide du logiciel Le Select. L'utilisateur peut manipuler cette vue uniforme des données au moyen de requêtes exprimées dans le langage Quilt. Les requêtes Quilt sont traduites en requêtes SQL exprimées sur un schéma relationnel générique. Toute donnée en format relationnel ou XML peut se décrire comme une vue (au sens base de données) exprimée sur ce schéma générique. Une étape de réecriture transforme la requête SQL exprimée sur le schéma générique en une requête exécutable par Le Select sur les sources de données concernées. Les données résultant de cette exécution sont ensuite traduites en format XML et retournées à l'utilisateur. L'intérêt majeur d'Agora est de permettre l'interrogation efficace de données relationnelles et XML dans une même requête Quilt. Ce logiciel a été démontré à la conférence VLDB'2000 avec Le Select mais ne fait pour l'instant pas l'objet de diffusion extérieure au projet.

5.3 Ajax

Participants : Helena Galhardas [correspondant], Eric Simon, Cristian Saita.

Le logiciel AJAX est un outil d'aide à la génération de programmes efficaces pour le nettoyage de données. AJAX offre un langage de haut niveau pour la spécification de programmes de nettoyage de données. Un programme dans ce langage décrit un graphe à flôts de données dont les noeuds sont des opérations de nettoyage. AJAX propose un ensemble de cinq opérateurs logiques paramétrables qui peuvent exprimer toutes les opérations de transformation de données nécessaires. AJAX offre également un environnement de mise au point sophistiqué qui permet d'inspecter le déroulement d'un programme et d'intervenir manuellement sur le résultat des transformations, de solliciter explicitement l'assistance de l'utilisateur depuis le programme de nettoyage via la génération d'exceptions et d'assister l'utilisateur dans le deboguage d'un programme via un mécanisme d'explication des exceptions générés. AJAX genère du code Java qui optimise l'exécution des opérations logiques de transformation. Le prototype a été présenté lors de la conférence SIGMOD'2000. Il est actuellement utilisé afin de nettoyer les 2 millions de références bibliographiques en informatique du site Web Citeseer qui sont collectées automatiquement sur le Web. Les premières expériences réalisées pour 100.000 références ont montré la puissance du langage de spécification d'AJAX et l'intérêt des techniques d'optimisation mises en oeuvre.

5.4 Le Subscribe

Participants : Françoise Fabret, François Llirbat [correspondant], Joao Pereira,

Alexandru Carstoiu.

Le Subscribe est un système de publication/souscription (« publish/subscribe », en anglais) dont le développement a débuté en 1999. Ce système est dédié à la diffusion en mode « push » d'informations ayant la forme d'ensembles de couples « attribut-valeur » appelés événements. Le langage de souscription supporté par le système est simple : chaque souscription consiste en une conjonction de prédicats sur les valeurs des attributs. L'objectif principal de ce système est de supporter un très grand nombre de souscriptions (plusieurs millions) et un haut débit d'événements (plusieurs centaines par seconde). Les souscripteurs et les éditeurs peuvent communiquer avec le système en utilisant le protocole Java RMI ou HTTP. Le Subscribe est composé de plusieurs composants logiciels, chacun étant responsable d'une fonctionnalité du système : filtrage des événements, notification des événements auprès des souscripteurs, etc, .. Ces composants peuvent être répartis sur plusieurs machines ou résider sur une même machine. Le système offre différents modes de notification : par e-mail, de façon immédiate en utilisant le protocole UDP ou Java RMI, ou sur demande des souscripteurs. Le point fort du système réside dans son module de filtrage qui implémente des algorithmes de pattern matching très performants. Le Subscribe a été démontré au Caire lors de la conférence VLDB'2000. Il a aussi été démontré en Janvier 2001 à New-York dans le cadre de la conférence et de l'exposition Linuxworld.

Cette année, les algorithmes de filtrage de Le Subscribe ont été étendus pour filtrer les documents XML. Dans ce contexte, le langage de souscription est un sous-ensemble du langage X-Path qui est un langage standard d'intérrogation des document XML. Cet algorithme a été démontré à Rome lors de la conférence VLDB'2001.

5.5 Weave

Participant : Khaled Yagoub [correspondant].

Le logiciel Weave est un système de gestion de sites Web à usage intensif de données. Il permet de construire des sites de façon déclarative ce qui facilite leur conception et leur mise en oeuvre et réduit le coût de leur maintenance. Weave possède une architecture configurable de caches à plusieurs niveaux permettant de cacher des données extraites d'une base de données sous forme de tables relationnelles, de fragments XML ou de pages HTML. La possibilité de cacher des fragments de données XML assure un contrôle sémantique très fin des informations cachées ce qui est très important dans des sites manipulant des données avec des droits d'utilisation limités (e.g., oeuvres d'art). Weave offre un langage de spécification déclaratif (appelé WeaveL) qui permet de définir le schéma d'un site (c'est-à dire de sa structure en pages et en hyper-liens), et la spécification de différentes stratégies de caching. Weave offre également un environnment complet de suivi des performances d'un site Web via la génération de statistiques sur l'utilisation du site. Weave a été démontré dans plusieurs conférences internationales (EDBT'00, WWW'00 et VLDB'00). Il est utilisé pour la construction et la gestion des sites Web des projets Caravel à l'Inria-Rocquencourt (`http://www-caravel.inria.fr`) et Aida à l'IRISA (`http://www.irisa.fr/aida/aida-new`).

5.6 Attman

Participants : Alberto Lerner, Eric Simon [correspondant].

Le logiciel Attman, développé en collaboration étroite avec Dennis Shasha (NYU, USA), offre un mode de navigation non hiérarchique dans des collections de données. Ce système utilise un modèle de données original composé de tables relationnelles et de cinq types de dépendances qui expriment des liens sémantiques entre les données. Chaque dépendance définit une relation de pertinence entre les données : les données d'une table sont pertinentes si elles satisfont les relations de pertinence auxquelles elles participent. L'utilisateur qui se connecte au système voit une liste de tables dont le contenu est consultable. Puis l'utilisateur peut sélectionner des lignes pertinentes dans une table. Le système calcule alors toutes les données qui demeurent pertinentes au regard de cette sélection en utilisant les dépendances définies par le concepteur de l'application. La liste des tables pertinentes restantes est ensuite présentée à l'utilisateur. Chacune de ces tables ne contient que des lignes pertinentes. L'utilisateur peut alors continuer sa navigation. En dehors du modèle de données, nous avons développé des algorithmes efficaces qui permettent d'effectuer le calcul des données pertinentes. Le système possède une architecture à trois tiers qui permet à une application cliente de se connecter à un serveur Attman qui puise ses données dans un serveur de données. Le système est opérationnel et fait l'objet d'une expérimentation afin de construire un site Web adaptatif pour JavaDoc.

6 Résultats nouveaux

La présentation des résultats de recherche est organisée selon les trois thèmes de recherche du projet présentés en Section 2. Globalement, les contributions ont surtout porté cette année sur la conception d'algorithmes et de méthodes d'optimisation et leur validation au travers de nos composants logiciels. On ne présente que les actions de recherche ayant donné lieu à des publications au cours de l'année 2001.

6.1 Accès à des ressources distribuées

Mots clés : hétérogénéité, optimisation, répartition.

Résumé : *le problème général abordé dans ce thème est celui de la publication de ressources dans un réseau ainsi que l'accès à ces ressources au moyen de langages de haut niveau. Les ressources peuvent être des données (structurées ou non) ou des services (librairies, programmes scientifiques, sites Web, etc), l'ensemble formant un* système d'information global. *Le système de médiation a la charge de mettre en relation de façon transparente les éditeurs des ressources avec les clients qui veulent utiliser ces ressources. Les actions de recherche qui structurent ce thème se distinguent selon l'approche, push ou pull, utilisée. Pour l'approche pull nous nous sommes concentrés sur deux questions essentielles. La première question est quelle vue uniforme des données faut-il présenter à l'utilisateur et quel est le langage de requêtes associé. Jusqu'à présent nous avons centré nos recherche sur*

le modèle relationnel et le langage SQL. Cette année, nous avons mené des recherches sur l'utilisation de XML comme formalisme de représentation uniforme des données, et sur la conception d'un langage de requêtes pour XML. La deuxième question est quelles techniques d'exécution efficaces peut-on élaborer pour les langages de requêtes supportés par le système de médiation. Dans l'approche push les clients sont intéressés par des informations hautement volatiles et dynamiques (des événements) ; ils souscrivent des abonnements auprès du médiateur indiquant les informations qui les intéressent, les éditeurs font parvenir leurs informations (événements) au médiateur. Ce dernier se charge d'alerter les souscripteurs chaque fois qu'une information intéressante est émise par un éditeur. Dans la mesure où l'utilisation de médiateurs push dans les applications web telles les applications de commerce électronique « B2C » (bourse d'échange, billeterie, informations sur le trafic,...) est conditionnée par la faculté du système à supporter un grand nombre de clients et un flux élevé d'événements, la question est quelles sont les techniques de filtrage efficaces dans un tel contexte.

6.1.1 Intégration de ressources hétérogènes au moyen de patterns d'accès : méthodologie et optimisation.

Participants : Luc Bouganim, Francoise Fabret, Aurelian Lavric, Ioana Manolescu, Eric Simon.

Dans le cadre des applications environnementales que nous envisageons avec Le Select, les scientifiques peuvent typiquement avoir à poser des requêtes impliquant des données et des programmes (par exemple, un programme d'extraction de motifs dans une image) publiés en divers points du réseau. Un des buts poursuivis par Le Select est de faciliter au maximum la tâche de formulation des requêtes et celle de publication des ressources (données et programmes). La solution retenue est de choisir le format relationnel pour représenter les ressources. Les ressources sont publiées au moyen de wrappers (extracteurs) sous forme de tables relationnelles à accès resteints. Par exemple, une fonction F de signature X -> Y est représentée par une table F(X,Y) avec restriction d'accès : on ne peut obtenir la valeur de l'attribut Y que si l'on fournit la valeur de X. Les restrictions d'accès aux données sont modélisées au moyen du concept de patterns d'accès (binding patterns). Cette modélisation des ressources est un point fort de Le Select : elle procure à la fois la simplicité de publication, la souplesse dans l'intégration et elle permet d'accéder aux ressources au moyen de requêtes formulées dans le langage standard SQL. Cependant bien que ces requêtes soient, à première vue, « classiques », leur exécution se distingue de celle de requêtes SQL standards car elle inclut l'appel itératif de programmes souvent très couteux en temps de calcul, elle conduit à manipuler (en particulier à transporter sur le réseau) des objets volumineux, elle est distribuée sur plusieurs serveurs. Ces particularités nécessitent d'élaborer des mécanismes d'exécution spécifiques. Nous proposons une solution à la fois algorithmique et architecturale. Pour exploiter efficacement les ressources à accès restreint, nous proposons un opérateur BindJoin qui supporte le caching et qui est parallélisable. Ces caractéristiques apportent une réponse particulièrement bien adaptée pour minimiser le surcoût induit par la présence de fonctions dans une requête. En effet, l'utilisation

de techniques de caching permet d'éviter d'exécuter plusieurs fois un même programme sur les même arguments. Nous proposons des algorithmes pour le BindJoin qui exploitent à la fois le cache et le parallélisme à l'intérieur même de l'opérateur de façon à produire la plupart des résultats au début de l'exécution. En ce qui concerne le transfert des objets volumineux, le but à atteindre est de ne transférer ce type d'objets que sur les sites où ils sont utiles, et à éviter de transférer plusieurs fois le même objet. Notre solution consiste à coupler les capacités de caching du BindJoin avec les services fournis par un composant logiciel chargé de la gestion de ce type d'objets dans chaque site.

6.1.2 Intégration de données XML et relationelles

Participant : Ioana Manolescu.

Dans certains types d'applications d'intégration de données très structurées (tables relationelles) et de documents (qui peuvent être modélisées en XML), il est nécessaire d'être capable de répondre à des requêtes portant sur les deux types de données en même temps. Un exemple de ce genre d'application est la gestion des données médicales : certaines informations font référence aux données personnelles des malades et sont plutot structurées, tandis que le dossier médical cummule les annotations faites par le personnel médical au cours d'une longue période de temps (parfois plusieurs années) et est plus proche par structure d'un document XML. Le médiateur Le Select, réalise déjà l'intégration de données relationelles ; nous l'avons donc étendu avec la capacité d'interroger, en même temps que des données relationelles, des documents XML. L'approche adoptée fonctionne sur quatre principes de base. Tout d'abord, nous avons defini un schema générique relationnel virtuel qui permet de représenter n'importe quel document XML. Le second principe est l'expression des sources de données XML comme des vues sur ce schema générique relationnel. Pour cela, un wrapper DOM (pour Le Select) a été construit afin de présenter l'interface standard DOM de manipulation d'un document XML sous la forme des tables du schema générique accompagnées des restrictions d'accès correspondantes. Chacune des tables exportées par le wrapper DOM peut ensuite être exprimée de façon naturelle comme une vue sur le schema générique virtuel. Le troisième principe est la traduction par défaut des tables exportées par les autres wrappers (non DOM) comme des documents XML. Dès lors, un utilisateur a la possibilite de voir les sources non XML et les sources XML comme des documents XML. Le dernier principe est une méthode de traduction progressive d'une requete XQuery en une requête SQL exécutable par les wrappers DOM et autres. Cette méthode de traduction constitue le point dur de l'approche. En particulier, notre travail a d'abord consisté à caractériser formellement le sous ensemble des requêtes XQuery qu'il était possible de traduire en requêtes SQL portant sur le schema générique virtuel sans matérialisation de résultats intermédiaires. Nous avons ensuite proposé une méthode de traduction en deux étapes qui consiste d'abord à normaliser une requête XQuery en une forme canonique puis á traduire cette forme canonique en SQL. Enfin, nous avons caracterisé parmi le sous ensemble precédent de requêtes XQuery celui qui produisait des requêtes SQL traduisibles en requêtes SQL portant sur les sources de données via une technique de réécriture de requêtes à l'aide de vues. Ces méthodes ont été implantées dans le logiciel Agora.

6.1.3 Algorithmes de pattern matching

Participants : Françoise Fabret, François Llirbat, João Pereira.

Le paradigme publication/souscription est utilisé dans de nombreuses applications pour découpler en temps, espace et en sujets d'intérêt des publieurs qui veulent disséminer de l'information et des souscripteurs qui désirent recevoir sélectivement des informations répondant à leurs intérêts. Un système publication/souscription doit assurer les fonctionnalités suivantes : enregistrer les souscriptions, recevoir l'information à disséminer, filtrer le contenu de cette information pour calculer les souscriptions satisfaites, notifier les souscripteurs dont les souscriptions sont satisfaites. Chacune de ces actions peut être un goulot d'étranglement lorsque le nombre de souscriptions (et de souscripteurs) devient très important, lorsque (de plus) les souscripteurs sont géographiquement situés sur une multitude de sites et lorsque, en même temps, le taux d'arrivée des informations est élevé. Nous avons concentré notre recherche sur le problème du filtrage efficace des contenus pour la dissémination d'informations (on parle aussi d'évenements). Nous nous nous sommes intéressé à deux types de contenus et de langage de souscription associés : Les contenus présentés sous forme de collections de paires (attribut,valeur) filtrés par des souscriptions exprimées sous forme de conjonctions de prédicats (attribut, comparateur, valeur) et les documents XML filtrés par des souscriptions exprimées en XPath. En ce qui concerne les contenus et souscriptions à base d'attributs, nous proposons un algorithme de pattern matching très performant capable de supporter un flôt très élevé dévenements et un très grand nombre de souscriptions (de l'ordre du million). L'algorithme s'adapte automatiquement aux changements de comportement des souscripteurs et de focus des informations à disséminer. Du point de vue technique, il s'agit d'un algorithme main memory, basé sur l'utilisation d'index multi-dimensionnels pour regrouper les souscriptions en clusters : les souscriptions sont modélisées sous forme de rectangles multi-dimensionnels et les évenements sous forme de points. La formulation du problème devient « étant donné un point, trouver les rectangles contenant ce point ». L'accès aux données via des index multi-dimensionnels a fait l'objet de nombreuses études (par exemple pour l'accès aux données géographiques) ; cependant les solutions existantes ne s'appliquent pas dans notre cas, et ceci pour deux raisons. Tout d'abord elles ne supportent qu'un nombre limité de dimensions (au mieux une dizaine) ; ensuite elles supportent mal les modifications des données indexées. Dans notre cas le nombre d'attributs (et donc de dimensions) peut être très élevé. De plus, les souscriptions sont très volatiles. Nous avons mis au point des structures d'index multi-dimensionnels répondant à ces problèmes. Le regroupement en clusters est dynamique. D'une part, les souscriptions ne sont pas indexées selon toutes leurs dimensions et le choix des dimensions à considérer pour une souscription est basé sur le pouvoir de sélectivité de ses divers prédicats. D'autre part, la décision de découper un cluster existant en plusieurs clusters est guidée par une fonction de coût tenant compte de la taille du cluster et des caractéristiques du flôt d'événements. Notre technique de découpage d'un cluster n'entraine que des modifications locales très peu coûteuses de l'index. Cette façon de découper les clusters rend l'agorithme très réactif aux comportements instantanés des souscripteurs et aux changements dans les statistiques sur les évenements. Du point de vue de l'implémentation, nous nous sommes placés dans le contexte des calculateurs « grande mémoire », et proposons des structures de données et des algorithmes de manipulation

de ces structures adaptés à ce contexte.

En ce qui concerne le filtrage des documents XML par des souscriptions en XPath, notre algorithme est lui aussi basé sur l'indexation des souscriptions. Ici, chaque souscription est indexée par une seule de ses dimensions. Nous avons mis au point des algorithmes efficaces pour la vérification des souscriptions atteintes via l'index.

6.2 Production de données dérivées

Mots clés : hétérogénéité, optimisation, répartition.

Résumé : *Lorsqu'on recherche une information précise dans un système d'information global, cette information n'existe pas toujours à l'état brut dans une source d'information. Dans ce cas, l'information doit être construite sur mesure (c'est souvent le cas dans les applications d'aide à la décision). Deux cas de figure assez différents se produisent. Dans le premier cas, la donnée recherchée peut être obtenue par intégration, consolidation, ou restructuration de données existantes. Une exemple typique est la construction d'entrepôts de données (« data warehouse » en anglais). Un problème crucial dans ce cas est d'obtenir des données sans doubles, sans erreurs et sans incoherences. On appelle ce probleme le nettoyage de données (« data cleaning » en anglais). Dans l'autre cas, les données recherchées peuvent s'obtenir au moyen d'un programme publié dans le système d'information global. Par exemple, on recherche une prédiction de l'évolution d'une nappe de pétrole lors d'une marée noire ; cette information peut être calculée par un programme scientifique de modélisation de l'évolution de nappes. Mais l'exécution de ce programme peut à son tour nécessiter des données d'entrée qui n'existent pas de façon brute et qui doivent être elles aussi calculées. On obtient ainsi une chaîne de traitement dont les étapes de traitement correspondent à l'exécution de programmes publiés. Dans ces deux cas de figure, toute la difficulté est de simplifier la mise en oeuvre du calcul de ces données recherchées et de garantir que les données produites ont la qualité désirée. Les deux actions qui composent ce thème ciblent ce problème avec des approches adaptées aux deux cas de figure cités.*

6.2.1 Modèles, langages et algorithmes pour le nettoyage de données

Participants : Helena Galhardas, Cristian Saita, Eric Simon.

Le problème du nettoyage de données est bien connu dans le domaine des systèmes d'aide à la décision et des entrepôts de données où il constitue l'un des problèmes les plus difficiles à résoudre. De nombreux outils, appelés ETL ou « data cleansing », ont été développés pour répondre à cette difficulté. Cependant, dans le cas d'applications telles que la migration de données très faiblement structurées vers des données structurées ou l'intégration de données scientifiques dans des domaines pluri-disciplinaires (e.g., la santé ou l'environnement), les outils existant destinés à l'écriture de programmes de nettoyage de données sont très insuffisants. Le problème principal rencontré est la conception d'un programme qui modélise un graphe à flots de données capable de transformer les données d'origine en données correctes et cohérentes et

qui s'exécute efficacement sur de gros volumes de données. Ce problème résulte de deux lacunes dans les systèmes existant : (i) le manque de séparation claire entre la spécification logique des opérations de transformation de données nécessaires et leur implantation physique, et (ii) le manque de fonctionalités permettant d'assister l'utilisateur dans la mise au point de son programme de nettoyage. Les recherches effectuées dans cette action ont répondu à ces lacunes par la proposition d'un langage déclaratif de spécification de programmes de nettoyage, d'un modèle d'exécution logique pour les opérations de nettoyage de données et des algorithmes efficaces de mise en oeuvre de ces opérations. Le modèle d'exécution intègre quatre opérateurs spécifiques (mapping, matching, clustering, merging) qui permettent de décomposer un flot de données en plusieurs flots, de recomposer plusieurs flots par comparaison de la similitude de leurs données, de partitioner un flot en groupes, ou de fusionner un groupe de données en une donnée unique. Tous ces opérateurs sont paramétrables par des fonctions fournies par l'utilisateur (par exemple, des fonctions de calcul de similitude entre données). Le langage utilise une syntaxe proche de SQL et permet d'exprimer les opérations du modèle d'exécution de façon déclarative. Il permet également de spécifier les conditions dans lesquelles des exceptions doivent être générées et l'utilisateur doit être sollicité pour une intervention manuelle dans le processus de nettoyage. Un mécanisme sophistiqué de gestion d'exceptions et d'explication assiste l'utilisateur à mettre au point son programme. Un effort particulier a porté sur la mise au point d'un algorithme de jointure par similitude pour des chaines de caractères qui est basé sur l'utilisation de techniques de filtrage. Plusieurs méthodes de filtrage ont été identifiées dans le cas ou la fonction de similitude sur des chaines est exprimée par une fonction « d'edit distance ». Nous avons montré que l'utilisation d'une certaine combinaison de filtres apportait les meilleurs gains de performance. Enfin, le choix d'algorithmes efficaces pour exécuter les opérations est effectué par un optimiseur. Toutes ces propositions ont été implantées dans le logiciel AJAX et validées sur une application de nettoyage des références bibliographiques utilisées par le site Web Citeseer avec un échantillon de 100.000 références. Cette validation a permis de vérifier l'utilité du langage et des fonctionnalités de gestion d'exceptions dans la mise au point de programmes ainsi que la performance des algorithmes proposés.

6.2.2 Workflow scientifiques pour le GRID

Participants : François Llirbat, Eric Simon, Jean-Pierre Matsumoto.

L'objectif principal du GRID est la globalisation des collaborations scientifiques à l'échelle du Web. L'idée est de permettre la mise en commun et l'utilisation partagée sur le réseau GRID de l'ensemble des données, des programmes, des modèles scientifiques ainsi que des capacités de calculs et de stockage. Une première étape pour atteindre cet objectif est la mise au point de technologies qui permettent, grâce à des réseaux très haut débits et des protocoles adéquats, le transport de grands volumes de données et l'utilisation intensive et en parallèle des capacitès de calcul disponibles sur le réseau. Ce premier objectif fait déjà l'objet de nombreux travaux de recherche dans le monde entier. Cependant pour que les scientifiques puissent facilement profiter de ces technologies, il faut aussi concevoir des outils de haut niveau qui facilitent la mise au point et la gestion d'expériences scientifiques sur le GRID. L'enjeu est ici d'abstraire les scientifiques des problèmes liés à l'accés et l'utilisation synchronisée des ressources du GRID

en donnant l'illusion d'un monde uniforme et centralisé.

Les workflow scientifiques, dérivés des workflows de gestion, ont été introduits comme un moyen pratique pour spécifier des expériences scientifiques. En effet, ils fournissent un modèle permettant une description formelle des expériences facilitant ainsi leur ex'ecution automatisée. Ils permettent aussi l'archivage et l'interrogation des expériences passées. Cependant les solutions actuelles imposent une gestion centralisée des données et des expériences autour d'une même base de données. Elles s'adaptent donc mal au contexte de distribution grande échelle imposé par le GRID. Adapter ces solutions au contexte du GRID impose de résoudre plusieurs problèmes difficiles. Un premier problème est la publication des données et des programmes sur le réseau de façon à permettre leur utilisation par d'autres scientifiques. Ce problème de publication est un problème difficile car il faut fournir un environnement de publication dans lequel les scientifiques puissent spécifier les conditions (parfois complexes) dans lesquelles leurs données ou leurs programmes peuvent être utilisés. Une fois ces informations publiées, une deuxième difficulté est de fournir un langage de définition des expériences qui permette de faire abstraction des problèmes de distribution et d'hétérogénéité des données et des programmes. De plus ce language doit être simple pour utilisable par des non informaticiaens, et suffisamment puissant pour supporter des chaînes de traitement complexes. Les recherches que nous avons effectuées dans ce domaine ont conduit à la mise au point d'un nouveau modèle de publication des ressources (données et programme) et à un langage de définition de workflows scientifiques. Le modèle de publication est construit sur deux niveaux. Le premier niveau est constitué par l'information brute comprenant le code des programmes et le contenu des données directement utilisées par les programmes. Le deuxième niveau est constitué de l'information contextuelle qui décrit chaque donnée et chaque programme ainsi que les contraintes opérationnelles associées. Le langage de workflow permet de décrire de façon déclarative et non ambigüe le flôt de données entre les programmes ainsi que leur synchronisation. Ce nouveau language a deux qualités importantes. Tout d'abord il supporte la notion de calcul itératif et le le concept d'événemnent. Ces concepts sont particulièrement adaptés au contexte des expériences scientifiques qui consistent souvent en des itérations successives sur une séquence de données ou en des traitements à la volée de données dynamiques. Ensuite, l'aspect déclaratif de notre language rend transparente l'utilisation de techniques d'optimisations pour l'exécution efficace des workflows. Pour faciliter ces optimisations nous avons ainsi mis au point un modèle d'exécution à base d'évènements qui permet une exécution parallèle maximale des programmes impliqués dans le workflow.

6.3 Consultation de données pour « tous »

Résumé : *Dans ce thème nous abordons le problème de la présentation de données à des utilisateurs « naïfs », ce qui sous-entend que les utilisateurs ne sont pas capables d'exprimer des requêtes dans un langage de bases de données tel que SQL, OQL ou XQuery. Les sites Web sont des instruments appropriés pour cela, car ils proposent un mode conversationnel très simple, basé sur la navigation. Mais l'accès navigationnel à des bases de données par le web pose des problèmes de performances. Le thème comporte deux actions. La première action vise à développer un système qui facilite la gestion du contenu et de la structure de sites Web tout*

en garantissant de bonnes performances d'accès grâce à l'utilisation de techniques d'anté-mémorisation. La deuxième action vise à développer un système qui offre une alternative à la présentation hiérarchique d'informations – telle qu'on la rencontre par exemple dans les catalogues électroniques sur le Web.

6.3.1 Algorithmes pour interfaces à base de requêtes dynamiques

Participants : Alberto Lerner, Eric Simon.

Le modèle semi-structuré qui est à la base de la structure des sites Web n'est pas toujours un bon support de navigation dans les données. Cette inadaptation se revèle lorsqu'un site renfermant de grandes collections de types de données est accédé par des utilisateurs ayant des profils assez differents et frequentant occasionellement le site. Dans ce cas, il est difficile d'anticiper la meilleure façon de structurer les données sous la forme d'un graphe (le graphe induit de la structure XML du site) de telle sorte que les intentions de navigation des utilisateurs ne soient pas bridées ou ne deviennent trop complexes (i.e., il faut suivre beaucoup de liens pour arriver à ce qu'on cherche). Une alternative est le paradigme des interfaces a base de requêtes dynamiques. Dans ce contexte, une requête permet de retenir ou d'exclure des données parmi un panel de données qui sont présentées à l'utilisateur. Une session utilisateur consiste en une suite de requêtes créées dynamiquement. L'efficacité se mesure alors en nombre de requêtes nécessaires à l'utilisateur pour parvenir à trouver les données qu'il cherche. Différentes interfaces graphiques (DQI « Dynamic Query Interface » en anglais) ont été conçues en suivant ce paradigme. Ces interfaces offrent le moyen de sélectionner en peu de requêtes les données d'une table relationnelle qui possède éventuellement beaucoup d'attributs. Le principe est le suivant. A chaque étape, l'utilisateur a le moyen de retenir ou d'exclure des données selon un critère exprimé sur un seul attribut à la fois (e.g., au moyen de curseurs, de menus ou de zooms sur une carte bi-dimensionnelle) ; puis le système lui affiche d'une façon particulière les données restantes. L'utilisateur consulte le résultat et formule un autre critère qui peut être soit plus restrictif qu'un critère précédent soit moins restrictif. Les études ergonomiques ont montré que ce type d'interface nécessitait de répondre à chaque critère en moins d'une seconde pour ne pas décourager l'utilisateur. Le problème est que les algorithmes existants garantissent cette contrainte de temps de réponse pour des volumes de données relativement faibles (quelque dizaines de milliers de tuples). Nous avons developpé un nouvel algorithme capable de répondre à une requête en moins d'une seconde pour des tables avoisinant le million de tuples. Un point important est que les techniques mises en oeuvre par notre algorithme sont compatibles avec les techniques d'optimisation déjà utilisées par les chercheurs dans la communauté des interfaces homme-machine.

7 Actions régionales, nationales et internationales

7.1 Actions régionales

A l'INRIA, nous entretenons depuis de nombreuses années une collaboration étroite avec le projet VERSO. Cette année, la collaboration a été marquée par la participation de François

Llirbat au projet Xylème. Nous coopérons aussi avec le projet AIR dans le domaine des systèmes d'information pour l'environnement, notamment pour les contrats européens THETIS et DECAIR. Enfin, nous avons collaboré avec le laboratoire PrIsm de l'Université de Versailles sur les problèmes d'accès à l'information distribuée et les workflow scientifiques.

7.2 Actions européennes

7.2.1 Environnement et climat DECAIR

L'objectif du projet DECAIR est de fournir des données de meilleure qualité aux organismes en charge de la prévision de la pollution urbaine. En particulier le projet se concentre sur la qualité des données fournies comme données d'entrée aux modèles de pollution de l'air. Ces données sont de différents types : données géographiques, données d'occupation des sols, données météorologiques, données d'émission de polluants. Pour atteindre cet objectif des efforts de recherche sont prodigués dans deux directions complémentaires : D'abord le projet explore la possibilité d'utiliser des données satellites pour améliorer la précision et la fraîcheur des données d'entrées. L'objectif est ici de fournir des méthodes et des algorithmes de traitement d'images satellites qui sont adaptés au problème de la pollution de l'air. De plus le projet étudie la mise au point d'un système d'information adapté capable d'accéder, traiter, transformer et intégrer des données provenant de plusieurs sources distantes comme les satellites, les stations aux sols, des bases de données. Ce système a en charge la maintenance automatique de la fraîcheur et de la qualité des données utilisées par les modèles. Pour valider cette approche, nous construisons un prototype appelé « démonstrateur DECAIR » capable de gérer l'exécution de la chaîne de traitement, de l'acquisition des images satellitaires jusqu'à la présentation des paramètres d'entrée aux modèles de qualité de l'air. Ce prototype sera testé avec deux modèles de qualité de l'air, l'un mesurant la qualité de l'air sur Madrid, l'autre sur Berlin. L'architecture du prototype doit être suffisament flexible pour permettre, dans des developpements futurs, d'élargir l'ensemble des données d'entrée qui peuvent être accédées automatiquement, d'intégrer et d'utiliser facilement de nouveaux modèles, de faciliter l'application de ces modéles à de nouveaux sites, de détecter et prendre en compte les changements météorologiques rapides en cours de l'exécution des modèles. Les partenaires de ce projet sont : le GMD à Berlin, l'UPM à Madrid, le CLRC-RAL en Angleterre, le FORTH-ICS en Grèce, BULL en France et le SICE en Espagne.

7.3 Actions internationales

7.3.1 Europe

- Ecole Polytechnique de Bucarest avec qui nous avons signé un protocole d'accord l'année dernière. Dans ce cadre, nous avons accueilli cette année trois étudiants roumains pour un stage de 6 mois. Deux d'entre eux se sont inscrit en DEA à Paris et effectueront leur stage au sein de l'équipe. L'année dernière nous avions accueilli cinq stagiaires dans le cadre de cet accord. Deux d'entre eux sont resté à Paris pour y obtenir un DEA. Ils ont effectué leur stage de DEA dans l'équipe et commence actuellement une thèse au sein de l'équipe.

- Yannis Ioannidis (Université d'Athènes) et Timos Sellis (NTUA, Athènes) avec qui nous travaillons sur les workflow scientifiques.
- Donald Kossman (Université de Munich) avec qui nous travaillons sur des techniques d'optimisation des langages de requêtes pour XML.
- Université technique de Lisbonne avec laquelle nous avons 2 contrats de coopération financés par la « Coopération Technique et Technologique Ambassade de France-ICCTI ».

7.3.2 Amérique du Nord

- IBM, Almaden, Californie. Nous travaillons avec Chandra Mohan sur l'optimisation dynamique de langages de requêtes et avec Don Chamberlin sur la conception du langage Quilt.
- Bell Labs, New Jersey (Narain Gehani, Rick Hull). Cette année, les résultats du travail de recherche mené par François Llirbat et Eric Simon avec les chercheurs de l'équipe Vortex dirigée par Rick Hull ont abouti à un transfert industriel important au sein de Lucent Technologies.
- NYU, New York. Dennis Shasha, avec qui nous développons de fortes collaborations sur les projets Ajax, Attman et Le Subscribe, a séjourné dans notre équipe pendant une semaine, et Eric Simon a effectué un séjour d'une semaine à NYU.
- Université d'Alberta, Edmonton, Canada (Tamer Özsu).
- Université de Toronto. Nous avons un projet de recherche franco-canadien en collaboration avec Arno Jacobsen.

7.3.3 Amérique du Sud et Amérique Centrale

- universités de Rio de Janeiro (PUC, UFRJ IME et UNI-Rio), avec lesquelles nous avons un projet de coopération CNPQ-Inria (ECOBASE) sur les systèmes d'information pour l'environnement. Nous avons organisé conjointement un workshop international sur l'intégration d'information à Rio en avril 2001 (WIIW'2001).

8 Diffusion de résultats

8.1 Animation de la Communauté scientifique

L'équipe a participé aux comités de programme des colloques suivants :
- Int. Conf. on Very Large Databases (VLDB) : M. Bouzeghoub.
- ACM SIGMOD Conference : E.Simon.
- Conf. Nationale BDA : E. Simon.
- Int. Conf. on Data Engineering (ICDE) : F. Llirbat.

L'équipe contribue aussi à des comités de lecture et associations :
- Int. Journal on Distributed and Parallel Database Systems, Kluwer Academic Publishers (E. Simon).
- Network and Information Systems Journal, Hermes (M. Bouzeghoub, rédacteur en chef, E. Simon).

8.2 Enseignement

Eric Simon occupe un poste de Directeur-Professeur à temps partiel dans le département de Génie Informatique de l'Ecole Supérieure d'Ingénierie Léonard de Vinci depuis le 1er Avril 2000.

- Entrepôts de données, PULV, 40 heures : E. Simon.
- Cours d'algorithmique, PULV, 10 heures : F. Llirbat.
- Bases de données, PULV, 20 heures : F. Llirbat.
- Bases de données à objets, Université Paris Sud (MIAGE 2ème année), 6 heures (I. Manolescu).
- Travaux dirigés en bases se données, algorithmique et structures de données, PULV : J-P. Matsumoto, M. Amzal, A. Lavric.

9 Bibliographie

Thèses et habilitations à diriger des recherches

[1] H. Galhardas, *Nettoyage de Données Déclaratif : Langage, Modèle, et Algorithmes*, thèse de doctorat, September 2001.

[2] I. Manolescu, *Optimization Techniques for Querying Heterogeneous Distributed Data Sources*, thèse de doctorat, December 2001.

[3] F. Porto, *Strategies for parallel execution of queries in distributed scientific dabase*, thèse de doctorat, 2001.

[4] K. Yagoub, *Data-Intensive Web Sites Specification and Optimization*, thèse de doctorat, May 2001.

Articles et chapitres de livre

[5] L. Bouganim, M.C. Cavalcanti, F. Fabret, M. andF. Llirbat, M. Mattoso, R. Melo, A.M. Moura, E. Pacitti, F. Porto, M. Simoes, E. Simon, A. Tanaka, P. Valduriez, « The Ecobase Project : Database and Web Technologies for Environmental Information Systems », *ACM SIGMOD Record*, september 2001.

[6] L. Bouganim, M.L. Campos, M.C. Cavalcanti, F. Fabret, F. Llirbat, M. Mattoso, R. Melo, A.M. Moura, E. Pacitti, F. Porto, M. Simões, E. Simon, A. Tanaka, P. Valduriez, « The Ecobase Environmental Information System : Applications, architecture and open issues », *Network and Information Systems Journal, Hermes, NISJ, Vol.3, No.5*, 2000.

[7] F. Fabret, D. Florescu, K. Yagoub, *Bases de données et internet : Modèles, langages et système*, Hermes, 2001, ch. 11.

[8] F. Llirbat, J. Matsumoto, E. Simon, J.P. Berroir, I. Herlin, H. Yahia, « Using Scientific Workflow Techniques for Automatic Processing Of environmental Data », *Journal of Systems Analysis Modelling Simulation (SAMS,to appear).*

[9] M.J. Blin, F. Fabret, « A cooperative Work Framework Integrating Collaboration and Coordination », *Cahiers du Lamsade no.179*, April 2001.

[10] P. Pucheral, L. Bouganim, P. Valduriez, C. Bobineau, « PicoDBMS : Scaling down Database Techniques for the Smartcard », *VLDB Journal (to appear)*, 2001.

Communications à des congrès, colloques, etc.

[11] N. Anciaux, C. Bobineau, L. Bouganim, P. Pucheral, P. Valduriez, « PicoDBMS : validation and experience », *in : Proc. of Int. Conference on Very Large Databases, VLDB*, Rome, september 2001.

[12] L. Bouganim, F. Fabret, F. Porto, P. Valduriez, « Exécution de requêtes comprenant des fonctions chères et des objets volumineux dans une architecture médiateur distribuée », *in : Proc. of Conférence Bases de Données Avancées, BDA'01*, Agadir, Maroc, Octobre 2001.

[13] L. Bouganim, F. Fabret, F. Porto, P. Valduriez, « Processing Queries with Expensive Functions and Large Objects in Distributed Mediator Systems », *in : Int. Conference on Data Engineering, ICDE 2001*, Heidelberg, Germany, April 2001.

[14] F. Fabret, F. Llirbat, J.Pereira, K. Ross, D. Shasha, A. Jacobsen, « Filtering Algorithms and Implementation for Very Fast Publish/Subscribe Systems », *in : Proc. of ACM SIGMOD Conf. on Management of Data*, Santa Barbara, USA, 2001.

[15] H. Galhardas, D. Florescu, D. Shasha, E. Simon, C. Saita, « Declarative Data Cleaning : Language, Model and Algorithms », *in : Proc. of Int. Conference on Very Large Databases, VLDB*, Rome, september 2001.

[16] H. Galhardas, D. Florescu, D. Shasha, E. Simon, C. Saita, « Improving Data Quality using a Data Lineage Facility », *in : Proc. of the Int. Workshop on Design and Management of Data Warehouses, DMDW, S. Gatziu et al (eds), workshop paper*, Heidelberg, june 2001.

[17] F. Llirbat, A. Jacobsen, « Publish/Subscribe Systems », *in : Tutorial in ICDE 2001,International Conference on Data Engineering 2001*, Heidelberg, Germany, April 2001.

[18] I. Manolescu, D. Florescu, D. Kossmann, « Answering XML Queries over Heterogeneous Data Sources », *in : Int. Conference on Very Large Databases*, Rome, Septembre 2001.

[19] I. Manolescu, D. Florescu, D. Kossmann, « Answering XML Queries over Heterogeneous Data Sources », *in : Proc. of Conférence Bases de Données Avancées, BDA'01*, Agadir, Maroc, Octobre 2001.

[20] M.J. Blin, F. Fabret, O. Kapitskaia, F. Llirbat, « ProjectLeader : a Constraint-Based Support for the Distributed Design of Component-Based Products », *in : International Wokshop on Product Family Engineering PFE-4*, Bilbao, Spain, October 3-5 2001.

[21] J. Pereira, F. Fabret, A. Jacobsen, F. Llirbat, D. Shasha, « WebFilter : A High-throughput XML-based Publish and Subscribe System, demo paper », *in : Proc. of VLDB*, Rome, Italie, 2001.

[22] J. Pereira, F. Llirbat, F. Fabret, K. Ross, D. Shasha, R. Preotiuc-Pietro, « Publish/Subscribe on the Web at Extreme Speed », *in : LinuxWorld Conference*, New York, 2001.

Rapports de recherche et publications internes

[23] H. Galhardas, D. Florescu, D. Shasha, E. Simon, C. Saita, « Declarative Data Cleaning : Language, Model and Algorithms », *Rapport Technique n°4149*, Inria-Rocquencourt, march 2001, `http://www.inria.fr/rrrt/rr-4149.html`.

[24] I. Manolescu, L. Bouganim, F. Fabret, E. Simon, « Efficient Data and Program Integration Using Binding Patterns », *Rapport de Recherche n°4239*, Inria, Aout 2001, `http://www.inria.fr/rrrt/rr-4239.html`.

[25] I. Manolescu, D. Florescu, D. Kossmann, « Pushing XML Queries inside Relational Databases », *Rapport de Recherche n°4112*, Inria, Janvier 2001, `http://www.inria.fr/rrrt/rr-4112.html`.